Hierarchical Attention Networks for Different Types of Documents with Smaller Size of Datasets

Hon-Sang Cheong[1(✉)], Wun-She Yap[1], Yee-Kai Tee[1], and Wai-Kong Lee[2]

[1] Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Sungai Long, Malaysia
dexter6855@hotmail.com, {yapws,teeyk}@utar.edu.my
[2] Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Sungai Long, Malaysia
wklee@utar.edu.my

Abstract. The goal of document classification is to automatically assign one or more categories to a document by understanding the content of a document. Much research has been devoted to improve the accuracy of document classification over different types of documents, e.g., review, question, article and snippet. Recently, a method to model each document as a multivariate Gaussian distribution based on the distributed representations of its words has been proposed. The similarity between two documents is then measured based on the similarity of their distributions without taking into consideration its contextual information. In this work, a hierarchical attention network (HAN) which can classify a document using the contextual information by aggregating important words into sentence vectors and the important sentence vectors into document vectors for the classification was tested on four publicly available datasets (TREC, Reuter, Snippet and Amazon). The results showed that HAN which can pick up important words and sentences in the contextual information outperformed the Gaussian based approach in classifying the four public datasets consisting of questions, articles, reviews and snippets.

Keywords: Document classification · Machine learning · Hierarchical attention network · Accuracy · Dataset

1 Introduction

Document classification is one of the research area in natural language processing. The goal of document classification is to automatically assign one or more categories to a document by understanding the content of a document. Due to the massive usage of cloud storages, data analytics tools have been incorporated by data storage vendors into

Supported by the Collaborative Agreement with NextLabs (Malaysia) Sdn Bhd (Project title: Advanced and Context-Aware Text/Media Analytics for Data Classification).

J.-H. Kim et al. (Eds.): RiTA 2018, CCIS 1015, pp. 28–41, 2019.
https://doi.org/10.1007/978-981-13-7780-8_3

their products. The availability of the large amount of data have motivated the development of efficient and effective document classification techniques. Such document classification techniques found their applications in topic labelling [24], sentiment classification [16], short-text categorization [7] and spam detection [11].

Words in a document can be represented as embeddings that act as feature for document classification. A document with smaller size, also known as short text may not be recognized easily as compared to longer text due to the issue of data sparsity and ambiguity. Traditional approach to classify a short text is to represent texts with bag-of-word (BoW) vectors. Even though BoW is simple and straightforward, but it did not take consideration on the contextual information of the document.

To utilize more contextual information in a document or text, biterm topic model (BTM) [4] had been introduced. However, BTM model may suffer from curse of dimensionality problem due to the use of sparse vector representation. In recent years, much research has been devoted to tackle the curse of dimensionality by learning distributed representation of words in documents. A neural probabilistic language model was proposed by Bengio et al. [2] to learn a distributed representation for words that capture neighboring sentences semantically. Instead of using a neural probabilistic language model, Mnih and Kavukcuoglu [17] used training log-bilinear models with noise-contrastive estimation too learn word embeddings. They also found that the embeddings learned by the simpler models can perform at least as well as those learned by the more complex one. Along this direction, Pennington, Socher and Manning [19] proposed the use of global log-bilinear regression model that efficiently leverages statistical information by training only on the nonzero elements in a word-word co-occurrence matrix.

On the other hand, deep learning approaches had been proposed for document classification. These approaches include convolutional neural network (CNN) [10] and recurrent neural network (RNN) based on long short term memory (LSTM) [8]. Even though neural network based text classification approaches had been proved to be quite effective by Kim [10] and Tang et al. [23] independently, the annotation of each word in these approaches only summarizes the preceding words, but never consider the following words. Hence, Bahdanau et al. [1] proposed the use of a bi-directional RNN that considers both the preceding and following words.

Recently, Yang et al. [25] proposed a new approach based on deep learning. They named the new approach as hierarchical attention network (HAN). The intuition of the proposed model is simple where a document can be split into sentences and each sentence can be split into words. Thus, the HAN model was designed to capture these two levels that form a document. Instead of using bi-directional RNN, HAN used bi-directional gated recurrent unit (GRU). GRU is a new type of hidden unit proposed by Cho et al. [5]. GRU is inspired by the LSTM unit with a simpler structure that can be implemented easily. First, the preceding and following words in each sentence are considered by HAN model such that the more important words will be given higher weightage. Subsequently, the preceding and following sentences in a document are

considered such that the more important sentence will be given higher weightage. Finally, a document is classified based on such weightages. The efficiency of HAN had been proved against six publicly available datasets (Yelp 2013 [23], Yelp 2014 [23], Yelp 2015 [23], Yahoo Answer [26], IMDB Review [6] and Amazon [26]). Besides, Poon et al. [21] also demonstrated that HAN model is suitable for document level polarity classification.

All six datasets tested with HAN share the same similarities: (1) All documents are of the same type – user review; (2) Total number of documents in each dataset is large ranging from 335,018 to 3,650,000; (3) The size of vocabulary in each dataset is large ranging from 211,245 to 1,919,336. As short text classification is different with normal text classification due to the issue of data sparsity and ambiguity [7], thus the applicability of HAN on short text classification remains unclear.

Recently, Gaussian model proposed by Rousseau et al. [22] had demonstrated its effectiveness in recognizing short texts. The Gaussian method models each document as a multivariate Gaussian distribution based on the distributed representations of its words. The similarity between two documents is then measured based on the similarity of their distributions without taking into consideration its contextual information. However, contextual information is not taken into consideration in the proposed Gaussian model. To exploit the contextual information, a short text categorization strategy based on abundant representation was proposed by Gu et al. which subsequently outperformed the Gaussian model over a public dataset, Snippet [20].

In this paper, our contributions are listed as follows:

1. The efficiency of HAN model against different types of documents with smaller datasets and vocabulary is investigated.
2. The accuracy results of HAN against four selected datasets that consist of different types of documents are slightly better than state-of-the-art document classification methods.

2 The Hierarchical Attention Network Proposed by Yang et al.

A hierarchical attention network (HAN) was proposed by Yang et al. [25] for document classification with two unique characteristics: (1) It has a hierarchical structure that mirrors the hierarchical structure of documents; (2) it has two levels of attention mechanisms that applied at the word- and sentence-level to capture qualitative information when classifying a document. Figure 1 shows the architecture of HAN.

Assume that a document has n sentences s_i and each sentence contains m words. Let w_{it} with $t = 1, \ldots, m$ denotes the t-th word in the i-th sentence. HAN consists of the following components:

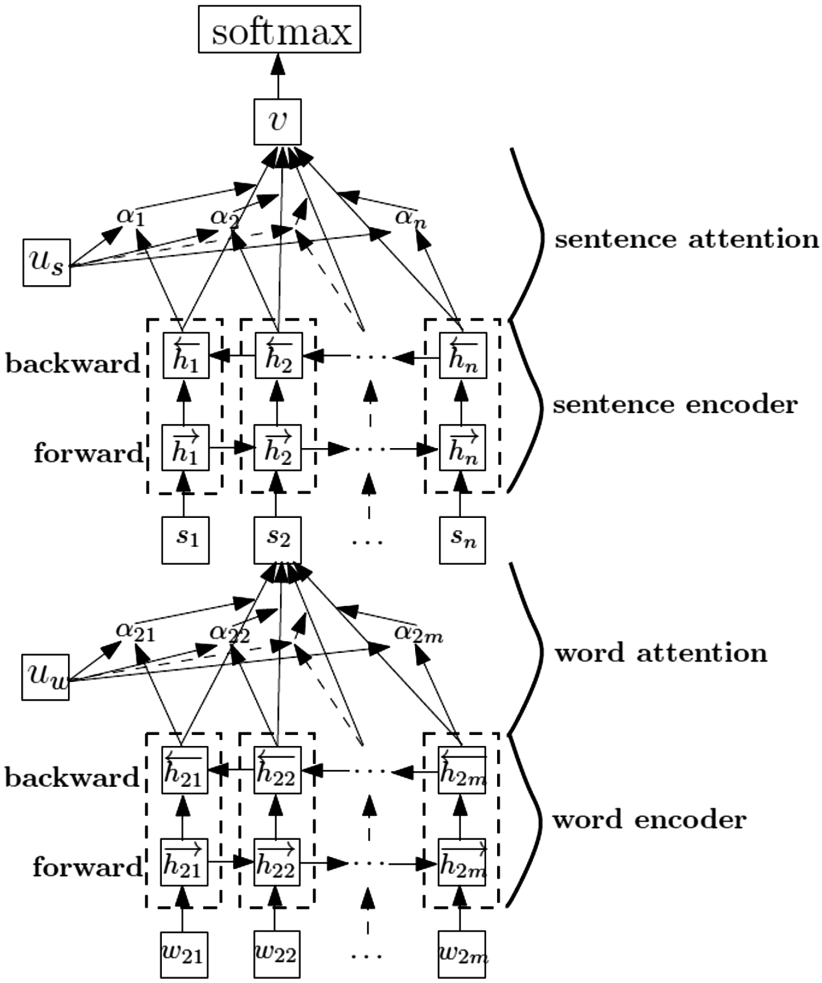

Fig. 1. The architecture of HAN proposed by Yang et al. [25]

1. **Gated Recurrent Unit (GRU) Based Sequence Encoder:** Cho et al. [5] proposed a new type of hidden unit, namely GRU that is inspired by the LSTM unit [8]. Differ with LSTM that has a memory cell and four gating units, GRU consists of two gating units only leading to simpler implementation and computation. The two gating units are named as reset gate r_t and update gate z_t for t-th hidden unit. These two gating units adaptively control the information ow inside the unit. First, the reset gate r_t is computed as follows:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{1}$$

where σ is the logistic sigmoid function, x_t is the input at time t, W_r and U_r are weight matrices which are learned and h_{t-1} is the previous hidden state. Then, the update gate z_t is computed as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{2}$$

where W_z and U_z are weight matrices which are learned. Subsequently, the candidate state $\widetilde{h}_t$ is computed as follows:

$$\widetilde{h}_t = tanh(W_h x_t + r_t \odot U_h h_{t-1} + b_h) \tag{3}$$

where W_h and U_h are weight matrices which are learned and $\odot$ is the element-wise multiplication. Notice that r_t controls how much the previous state contributes to the candidate state. Finally, the new state h_t is computed as follows:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t \tag{4}$$

2. **Word Encoder:** Given w_{it}, the words are embedded to vectors $x_{it} = W_e w_{it}$ where W_e is an embedding matrix. Bidirectional GRU [1] is then applied to get annotation of words by summarizing not only the preceding words, but also the following words. A bidirectional GRU consists of forward and backward GRU's, denoted as $\overrightarrow{GRU}$ and $\overleftarrow{GRU}$ respectively. The forward GRU reads the input sequence s_i as it is ordered from x_{i1} to x_{im} to calculate a sequence of forward hidden states $\overrightarrow{h}_{i1}, \ldots, \overrightarrow{h}_{im}$. Meanwhile, the backward GRU reads the input sequence as it is ordered from x_{im} to x_{i1} to calculate a sequence of forward hidden states $\overleftarrow{h}_{im}, \ldots, \overleftarrow{h}_{i1}$. The computations are listed as follows:

$$x_{it} = W_e w_{it},\ t \in [1, m] \tag{5}$$

$$\overrightarrow{h}_{it} = \overrightarrow{GRU}(x_{it}),\ t \in [1, m] \tag{6}$$

$$\overleftarrow{h}_{it} = \overleftarrow{GRU}(x_{it}),\ t \in [m, 1] \tag{7}$$

Finally, $h_{it} = \left[\overrightarrow{h}_{it}, \overleftarrow{h}_{it}\right]$ which summarizes the information of the whole sentence s_i centered around w_{it} is obtained.

3. **Word Attention:** Each word in a sentence s_i may not contribute equally to the representation of the meaning of a sentence. Thus, attention mechanism is included to extract and aggregate important words that contribute to the meaning of a sentence as a sentence vector as follows:

$$u_{it} = tanh(W_w h_{it} + b_w) \tag{8}$$

$$\alpha_{it} = \frac{\exp\left(u_{it}^{\top} u_w\right)}{\sum_i \exp(u_{it}^{\top} u_w)} \tag{9}$$

$$s_i = \sum_t \alpha_{it} h_{it} \tag{10}$$

where u_{it} is a hidden representation of h_{it} and u_w is randomly initialized and jointly learned during the training process.

4. **Sentence Encoder:** Given the sentence vector s_i, bidirectional GRU is applied to encode the sentences as follows:

$$\overrightarrow{h_i} = \overrightarrow{GRU}(s_i),\ i \in [1, n] \tag{11}$$

$$\overleftarrow{h_i} = \overleftarrow{GRU}(s_i),\ i \in [n, 1] \tag{12}$$

Finally, $h_i = \left[\overrightarrow{h_i}, \overleftarrow{h_i}\right]$ which summarizes the neighbour sentences around sentence i is obtained.

5. **Sentence Attention:** Each sentence may not contribute equally to the representation of the classification of a document. Thus, attention mechanism is included to extract and aggregate important sentences that contribute to the classification of a document as a document vector v as follows:

$$u_i = tanh(W_s h_i + b_s) \tag{13}$$

$$\alpha_i = \frac{\exp\left(u_i^\top u_s\right)}{\sum_i \exp(u_i^\top u_s)} \tag{14}$$

$$v_i = \sum_t \alpha_i h_i \tag{15}$$

where u_s is randomly initialized and jointly learned during the training process.

6. **Document Classification:** Document vector v can be used as features for document classification as follows:

$$p = \mathrm{softmax}(W_c v + b_c) \tag{16}$$

Finally, negative log likelihood of the correct labels is used as training loss as follows:

$$L = -\sum_d \log p_{dj} \tag{17}$$

where j is the label of document d.

3 Experiments

HAN was applied on six publicly available datasets by Yang et al. [25], and the results of HAN showed better accuracy as compared to the existing methods. These six publicly available datasets include the following:

1. Yelp'13, Yelp'14 and Yelp'15 [23]: Yelp reviews are obtained from the Yelp Dataset Challenge in 2013, 2014 and 2015. Ratings are given from 1 to 5 where higher rating is better.
2. IMDB reviews [6]: User ratings are given from 1 to 10 where higher rating is better.
3. Yahoo answers [26]: The document includes question titles, question contexts and best answers over 10 different classes.
4. Amazon reviews [26]: User ratings are given from 1 to 5 where higher rating is better.

Table 1 presents the summary of all these six datasets being tested using HAN by Yang et al. [25].

Table 1. Data statistics: #s denotes the number of sentences per document, #w denotes the number of words per document, word frequency is the ratio of # document to vocabulary [25]

	Yelp 2013	Yelp 2014	Yelp 2015	IMDB review	Yahoo answer	Amazon review
# Classes	5	5	5	10	10	5
# Documents	335,018	1,125,457	1,569,264	348,415	1,450,000	3,650,000
Average #s	8.9	9.2	9.0	14.0	6.4	4.9
Maximum #s	151	151	151	148	515	99
Average #w	151.6	156.9	151.9	325.6	108.4	91.9
Maximum #w	1184	1199	1199	2802	4002	596
Vocabulary	211,245	476,191	612,636	115,831	1,554,607	1,919,336
Word frequency	1.5859	2.3635	2.5615	3.0080	0.9327	1.9017

From Table 1, we observe the common similarities in the six publicly available datasets tested with HAN by Yang et al. [25]. These six datasets share the following similarities:

1. Each document is written by normal user consisting user's opinion toward certain topic.
2. Each document contains at least 4.9 sentences with 91.9 words in average.
3. Each dataset contains more than 200,000 vocabulary.
4. Each word appears in a dataset 0.9327 to 3.008 times in average.

Thus, experiments are conducted with the aim of answering the following research questions:

- **RQ1:** Do HAN proposed by Yang et al. [25] outperform the state-of-the-art methods in classifying different types of documents, e.g., questions, news article prepared by professional, brief description (snippet) and user review?
- **RQ2:** Do HAN proposed by Yang et al. [25] outperform the state-of-the-art methods in classifying the document which contains lesser training data, smaller vocabulary and/or lesser words?

3.1 Datasets - Selection

To answer **RQ1** and **RQ2**, the following four publicly available datasets are selected:

1. **TREC** [14]: Consists of a set of questions only (without user answers) that can be classified into six classes. These six classes are ABBREVIATION, DESCRIPTION, ENTITY, HUMAN, LOCATION and NUMERIC. One of the samples from the class DESCRIPTION is presented as follows for illustration purpose.

   title What is the oldest profession?

2. **Reuters-21578** [22]: Contains different business and financial news over more than 100 classes. Only eight classes with higher number of document per class are considered in this paper. These eight classes are EARN, ACQ, MONEY-FX, GRAIN, TRADE, INTEREST and SHIP. One of the samples from the class MONEY-FX is presented as follows for illustration purpose.

   ZAMBIA TO RETAIN CURRENCY AUCTION, SAYS KAUNDA Zambia will retain its foreign-exchange auction system despite the suspension of weekly auctions since January 24, President Kenneth Kaunda said.

3. **Amazon** [3]: Product reviewers acquired from Amazon over four different sub-collections, that is, BOOK, DVD, ELECTRONIC and KITCHEN. One of the samples from the class DVD is presented as follows for illustration purpose.

   I saw the scene,where they have Lissa chained to the pool table and gagged in the basement.I didn't understand most of the movie. I bet Kim Possible,Ron Stoppabl,and Rufus can deal with them

4. **Snippets** [20]: Contains word snippets collected from the Google search transactions that can be classified into eight classes. These eight classes are BUSINESS, COMPUTERS, CULTURE-ARTS-ENTERTAINMENT, EDUCATION-SCIENCE, ENGINEERING, HEALTH and SCIENCE. One of the samples from the class HEALTH is presented as follows for illustration purpose.

wikipedia wiki clinic clinic wikipedia encyclopedia clinic outpatient clinic public facility care ambulatory patients clients

Table 2 presents the summary of all selected four datasets. Different types of documents are included in these four datasets, that is, news article from Reuters, user review from Amazon, short description from Google snippets and question from TREC. The total documents for each selected dataset are at least 28 times smaller than to those datasets being tested by Yang et al. in [25] (see Table 1 for comparison). Similarly, the vocabulary for each selected dataset is at least five times smaller as compared of the datasets presented in Table 1. Thus, each word appears in the four selected datasets 0.2044 to 0.6257 times only in average.

Table 2. Data statistics: #s denotes the number of sentences per document, #w denotes the number of words per document, word frequency is the ratio of # document to vocabulary

	TREC	Reuters	Amazon	Snippets
# Classes	6	8	4	8
# Documents	5,952	7,528	8,000	12,340
Document type	Question	News	Review	Snippet
Average #s	1	6	7	1
Maximum #s	1	68	207	1
Average #w	10	138	128	17
Maximum #w	38	1,322	5,160	38
Vocabulary	9,513	23,582	39,133	29,276
Word frequency	0.6257	0.3192	0.2044	0.4215

3.2 Baseline

The following models are described and are included as baseline for performance comparison.

1. **BOW (binary)** [22]: All documents are pre-processed into bag-of-words vectors. If a word is present in the sentence, then its entry in the vector is 1; otherwise 0. Support vector machine (SVM) method is used to perform text classification.
2. **Centroid** [22]: Documents are projected in the word embedding space as the centroids of their words. Similarity of the documents is then computed using cosine similarity for text classification.
3. **NBSVM** [22]: Wang and Manning [24] combined both Naive Bayes classifier with SVM to achieve remarkable results on several tasks. Rousseau et al. [22] used a combination of both unigrams and bigrams as underlying features.
4. **WMD** [22]: Word Mover's Distance (WMD) is used to compute distances between documents [12]. Rousseau et al. [22] used pre-trained vectors from word2vec (i.e. a two-layer neural networks that are trained to learn linguistic

contexts of words from a large corpus of text) to compute distance between documents. Text classification is done with *k*-nearest neighbors (KNN) algorithm with the distances between documents. Notice that KNN algorithm classifies an object based on a majority vote of its *k* neighbors.

5. **CNN** [22]: CNN [13] exploits layer with convolving filters that are applied to local feature. Kim [10] showed that a simple CNN with little hyperparameter tuning and static vectors achieves excellent results for sentence-level classification tasks.
6. **DCNN** [9]: Dynamic *k*-max pooling is used on top of CNN for the semantic modelling of sentences.
7. **Gaussian** [22]: Short texts are treated as multivariate Gaussian distributions based on the distributed representations of its words. Subsequently, the similarity between two documents is then measured based on the similarity of their distributions for classification.
8. **DMM** [18]: Dirichlet Multinomial Mixture (DMM) model assumes that all documents are generated from a topic. Given the limited content of short texts, this assumption is reasonable.
9. **GPU-DMM** [15]: Inspired by DMM and the generalized Pólya urn (GPU) model, GPU-DMM was proposed by Li et al. [15] to promote the semantically related words under the same topic during the sampling process.
10. **BTM** [4]: Biterm Topic Model (BTM) learns the topics by modeling the generation of word co-occurrence patterns in short texts [4]. Biterm from BTM is an unordered word pair co-occurred from short context.
11. **Bi-RNN + Topic** [7]: Short texts are classified based on abundant representation which utilizes bi-directional recurrent neural network (CNN) with long short term memory (LSTM) and topic model to capture contextual and semantic information.

3.3 Experimental Settings and Results

Different common pre-processing techniques are performed on different selected datasets. These techniques include performing tokenization, removing stop word, removing special character, changing the capitalization of character and selecting pivots with mutual information. For our implemented HAN model, we use pre-trained word embedding vectors from global vectors for word representation (GloVe) to initialize the weight of word embedding layer. Notice that GloVe [19] is an unsupervised learning algorithm for obtaining vector representations for words. Different hyperparameters are set for different datasets as shown in Table 3. Notice that we split each document into a number of sentences denoted as # sentences.

Table 3. Different hyperparameter's settings for different selected datasets

Hyperparameter	TREC	Reuters	Amazon	Snippet
Word embedding dimension	100	200	100	300
GRU dimension	100	100	100	300
# sentences	1	1	10	1
# Training data	5,452	5,485	7,200	10,060
# Testing data	500	2,189	800	2,280

Table 4 shows the comparison of HAN [25] with aforementioned models on TREC, Reuters, Amazon and Snippet datasets in terms of the accuracy of document classification.

Regarding to **RQ1** and **RQ2**, as shown in Table 4, HAN which can pick up important words and sentences in the contextual information is able to out-perform all state-of-the-art models with the improvement of 0.28% to 0.78% in classifying the four public datasets that consists of different types of documents (i.e., question, review, news article and snippets), with smaller size of vocabulary, smaller training data and/or lesser words. This shows that HAN is also suitable for classifying documents with smaller size of vocabulary and lesser words.

Table 4. Comparison of different models for document classification in terms of accuracy

Method	TREC	Reuter	Amazon	Snippet
BoW (binary)	0.9660	0.9571	0.9126	0.6171
Centroid	0.9540	0.9676	0.9311	0.8123
NBSVM	0.9780	0.9712	0.9486	0.6474
WMD	0.9240	0.9502	0.9200	0.7417
CNN	0.9800	0.9707	0.9448	0.8478
DCNN	0.9300	–	–	–
Gaussian	0.9820	0.9712	0.9498	0.8224
DMM	–	–	–	0.8522
GPU-DMM	–	–	–	0.8722
BTM	–	–	–	0.8272
Bi-RNN + topic	0.9400	–	–	0.8636
HAN (this paper)	**0.9860**	**0.9790**	**0.9537**	**0.8750**

4 Visualization of Attention Mechanism

Yang et al. [25] showed that HAN is able to pick up important words and sentences for a user review which consists many words. In this section, we check whether HAN is able to pick up important words for a short question found from the class NUMERIC of TREC dataset. The raw question (without going through pre-processing) randomly selected from TREC is as follows:

dist How far is it from Denver to Aspen?

After going through pre-processing, the question mark is removed as follows:

dist How far is it from Denver to Aspen

Finally, Fig. 2 shows the visualization of attention mechanism for the selected question. Notice that the word with greater red color, the more important the word. This is done by first extracting out the word representation and subsequently coloring each word based on the word representation accordingly. Even though the question is short, HAN is still able to pick up important words that can classify the question as numeric such as "How". On the other hand, the word "is" is not so important for classification.

Fig. 2. The visualization of attention mechanism for the selected question from TREC (Color figure online)

5 Conclusion

In this paper, our results have demonstrated that HAN is suitable to classify different types of documents (review, question, snippet, and news article) with different sizes. We also showed that HAN is able to pick up important words even for question typed document. However, the improvement of accuracy in classifying short texts cannot be considered as significant. Thus, the future work includes the modification of HAN to further improve the accuracy in classifying both long and short texts.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv: 1409.0473 (2014)
2. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
3. Blitzer, J., Dredze, M., Pereira, F.: Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: Carroll, J.A., van den Bosch, A., Zaenen, A. (eds.) Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), pp. 440–447. Association for Computational Linguistics, Prague (2007)
4. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: topic modeling over short texts. IEEE Trans. Knowl. Data Eng. **26**(12), 2928–2941 (2014)
5. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1724–1734. Association for Computational Linguistics, Doha (2014)
6. Diao, Q., Qiu, M., Wu, C.-Y., Smola, A.J., Jiang, J., Wang, C.: Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In: Macskassy, S.A., Perlich, C., Leskovec, J., Wang, W., Ghani, R. (eds.) Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2014), pp. 193–202. ACM, New York (2014)

7. Gu, Y., et al.: An enhanced short text categorization model with deep abundant representation. World Wide Web **21**(6), 1705–1719 (2018)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1977)
9. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Toutanova, K., Wu, H. (eds.) Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), pp. 655–665. Association for Computational Linguistics, Baltimore (2014)
10. Kim, Y.: Convolutional neural networks for sentence classification. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1746–1761. Association for Computational Linguistics, Doha (2014)
11. Androutsopoulos, I., Koutsias, J., Chandrinos, K., Spyropoulos, C.D.: An experimental comparison of Naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In: Yannakoudakis, E.J., Belkin, N.J., Ingwersen, P., Leong, M.-K. (eds.) Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), pp. 160–167. ACM, Athens (2000)
12. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), pp. 957–966. Proceedings of Machine Learning Research, Lille (2015)
13. LeChun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
14. Li, X., Roth, D.: Learning question classifiers. In: Tseng, S.-C., Chen, T.-E. (eds.) Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), C02-1150. Howard International House and Academia Sinica, Taipei (2002)
15. Li, C., Wang, H., Zhang, Z., Sun, A., Ma, Z.: Topic modeling for short texts with auxiliary word embeddings. In: Perego, R., Sebastiani, F., Aslam, J.A., Ruthven, I., Zobel, J. (eds.) Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016), pp. 165–174. ACM, Pisa (2016)
16. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Lin, D., Matsumoto, Y., Mihalcea, R. (eds.) Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011), pp. 142–150. Association for Computational Linguistics, Portland (2011)
17. Mnih, A., Kavukcuoglu, K.: Learning word embeddings efficiently with noise-contrastive estimation. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Wein-berger, K.Q. (eds.) Proceedings of the Advances in Neural Information Processing Systems 26 (NIPS 2013), pp. 2265–2273. Neural Information Processing Systems Foundation, Lake Tahoe (2013)
18. Nigam, K., Mccallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Mach. Learn. **39**(2–3), 103–134 (2000)
19. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1532–1543. Association for Computational Linguistics, Doha (2014)
20. Phan, X.H., Nguyen, M.L., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Huai, J., et al. (eds.) Proceedings of the 17th International Conference on World Wide Web (WWW 2008), pp. 91–100. ACM, Beijing (2008)

21. Poon, H.-K., Yap, W.-S., Tee, Y.-K., Goi, B.-M., Lee, W.-K.: Document level polarity classification with attention gated recurrent unit. In: Knight, K., Nenkova, A., Rambow, O. (eds.) Proceedings of the 2018 International Conference on Information Networking (ICOIN 2018), pp. 7–12. IEEE, Chiang Mai (2018)
22. Rousseau, F., Vazirgiannis, M., Nikolentzos, G., Meladianos, P., Stavrakas, Y.: Multivariate Gaussian document representation from word embeddings for text categorization. In: Lapata, M., Blunsom, P., Koller, A. (eds.) Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017), vol. 1432, pp. 450–455. Association for Computational Linguistics, Valencia (2017)
23. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (eds.) Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), pp. 1422–1432. Association for Computational Linguistics, Lisbon (2015)
24. Wang, S.I., Manning, C.D.: Baselines and bigrams: simple, good sentiment and topic classification. In: Lin, C.-Y., Osborne, M. (eds.) Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), pp. 90–94. Association for Computational Linguistics, Jeju Island (2012)
25. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A.J., Hovy, E.H.: Hierarchical attention networks for document classification. In: Knight, K., Nenkova, A., Rambow, O. (eds.) Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016), pp. 1480–1489. Association for Computational Linguistics, San Diego (2016)
26. Zhang, X., Zhao, J.J., LeCun, Y.: Character-level convolutional networks for text classification. In: Cortes, C.A., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Proceedings of the Advances in Neural Information Processing Systems (NIPS 2015), pp. 649–657. Neural Information Processing Systems Foundation, Montreal (2015)